Data-driven lithium-ion battery capacity estimation from voltage relaxation

Jiangong Zhu
Tongji University
Yuan Huang
Tongji University
Michael Knapp
Institut für Angewandte Materialien
Xinhua Liu
Beihang University https://orcid.org/0000-0002-4111-7235
Yixiu Wang
University of British Columbia
R. Bhushan Gopaluni
University of British Columbia
Yankai Cao
University of British Columbia
Michael Heere
Karlsruhe Institute of Technology
Martin J Mühlbauer
Karlsruhe Institute of Technology
Liuda Mereacre
Karlsruhe Institute of Technology (KIT)
Haifeng Dai (tongjidai@tongji.edu.cn)
Tongji Unversity
Anatoliy Senyshyn
Technische Universität München https://orcid.org/0000-0002-1473-8992
Xuezhe Wei
Tongji University
Helmut Ehrenberg
Karlsruhe Institute of Technology (KIT) https://orcid.org/0000-0002-5134-7130

Article

Keywords: Lithium-ion battery, capacity estimation, voltage relaxation, data-driven method

Posted Date: August 13th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-770709/v1

Version of Record: A version of this preprint was published at Nature Communications on April 27th, 2022. See the published version at https://doi.org/10.1038/s41467-022-29837-w.

Data-driven lithium-ion battery capacity estimation from voltage relaxation

Jiangong Zhu [a, b], Yuan Huang [a,b], Michael Knapp [b *], Xinhua Liu [c], Yixiu Wang [a, d], R. Bhushan Gopaluni [d], Yankai Cao [d], Michael Heere [b,e], Martin J. Mühlbauer [b], Liuda Mereacre [b], Haifeng Dai [a *], Anatoliy Senyshyn [e], Xuezhe Wei [a], Helmut Ehrenberg [b]

[a] Clean Energy Automotive Engineering Center, School of Automotive Engineering, Tongji University, 201804 Shanghai, China
[b] Institute for Applied Materials (IAM), Karlsruhe Institute of Technology (KIT), 76344 Eggenstein-Leopoldshafen, Germany
[c] School of Transportation Science and Engineering, 100083, Beihang University, Beijing, China
[d] Department of Chemical and Biological Engineering, University of British Columbia, BC V6T 1Z3, Canada
[e] Heinz Maier-Leibnitz Zentrum (MLZ), Technische Universität München, Lichtenbergstr. 1, 85748 Garching b. München, Germany

Abstract:

Accurate capacity estimation is critical for reliable and safe operation of lithium-ion batteries. A proposed approach exploiting features from the relaxation voltage curve enables battery capacity estimation without requiring previous cycling information. Machine learning methods are used in the approach. A dataset including 27,330 data units are collected from batteries with $LiNi_{0.86}Co_{0.11}Al_{0.03}O_2$ cathode (NCA battery) cycled at different temperatures and currents until reaching about 71% of their nominal capacity. One data unit comprises three statistical features (variance, skewness, and maxima) derived from the relaxation voltage curve after fully charging and the following discharge capacity for verification. Models adopting machine learning methods, i.e., ElasticNet, XGBoost, Support Vector Regression (SVR), and Deep Neural Network (DNN), are compared to estimate the battery capacity. Both XGBoost and SVR methods show good predictive ability with 1.1 % root-mean-square error (RMSE). The DNN method presents a 1.5% RMSE higher than that obtained using ElasticNet and SVR. 30,312 data units are extracted from batteries with $LiNi_{0.83}Co_{0.11}Mn_{0.07}O_2$ cathode (NCM battery). The model trained by the NCA battery dataset is verified on the NCM battery dataset without changing model weights. The test RMSE is 3.1% for the XGBoost method and 1.8% RMSE for the DNN method, indicating the generalizability of the capacity estimation approach utilizing battery voltage relaxation.

Key words:

Lithium-ion battery; capacity estimation; voltage relaxation; data-driven method

Introduction

Lithium-ion batteries have become the dominant energy storage device for portable electric devices, electric vehicles (EVs), and many other applications [1]. However, battery degradation is an important concern in the use of lithium-ion battery as its performance decreases over time due to irreversible physical and chemical changes [2,3]. State of Health (SoH) has been used as an indicator of the state of the battery, and is usually expressed by the ratio of the relative residual capacity with respect to the initial capacity [4]. The accurate battery capacity estimation is challenging but critical to the reliable usage of lithium-ion battery, i.e., accurate capacity estimation allows an accurate driving range prediction and accurate calculation of the maximum energy storage capability in a vehicle. Typically, the battery capacity is gained by a full discharge process after it has been fully charged. In a real-life usage scenario, the battery full charge is often achieved while the EVs are parking with grid connection, however, the battery discharge depends on the user behavior with uncertainties in environmental and operational conditions, a complete discharge curve is seldom available for on-board battery health monitoring. The battery charging and discharging voltage, as one of the easily obtained parameters, depend on both, thermodynamic and kinetic characteristics of the battery. Thus, some capacity estimation methods using a partial charge/discharge process are proposed to estimate capacity for practical applications. The Ampere-hour counting method, which utilizes the accumulated charge value and the corresponding state of charge (SoC) change during a certain charge or discharge process is used to calculate the battery capacity [4]. A partial charge process in a specific voltage range as a feature is used for on-board capacity estimation [5]. Some transformations of the partial voltage curves, i.e., differential voltage analysis [6] and incremental capacity analysis [7,8], are used for battery capacity loss evaluation and aging mechanism identification. Eddahech et al. [9] proposed a battery SoH determination method based on a constant-voltage charge phase for batteries under calendar aging. A battery capacity estimation method which utilizes the resting process after charging or discharging, that is, the relaxation voltage process, is also proposed. Baghdadi et al.[10] used the relaxation voltage after full charge and 30 min of rest, and proposed a linear model to estimate battery capacity for three different commercial batteries. Schindler et al. [11] and C. Lüders et al. [12] took the voltage relaxation for the lithium plating detection in the battery capacity fade process. The relaxation process was also transferred to several models for parameterization [13-15]. Literature shows that the equivalent circuit model (ECM) provides excellent accuracy in terms of modeling the battery transient and steady state. The ECM is composed of multiple resistor-capacitor (RC) parallel links to accurately fit the relaxation voltage curve over a long period of time. Qian et al. [16] used a second-order ECM to describe the voltage relaxation, and found that the extracted parameters provided an evaluation of the battery SoH and aging mechanisms. Attidekou et al. [17] modelled the battery capacity decay during rest periods at 100% SoC using a dynamic time constant derived from the RC network model. Fang et al.[18] proposed a battery SoH estimation method based on the linear relationship between feature parameters and open circuit time during the battery relaxation even under different SoCs. However, as the amount of RC links increases, the complexity of the ECM will increase accordingly, which makes it difficult to use in an on-board application [19]. Besides, the accuracy and robustness of capacity estimation are difficult to evaluate because of the differences in battery types and working conditions.

The data-driven methods using statistical and machine learning techniques have been popular in battery research recently [20,21]. The data-driven methods do not need

deep understanding of battery electrochemical principles, but large numbers of data are required to ensure the model reliability [22]. Severson et al. [23] reported a promising route using machine learning to construct models that accurately predicted $LiFePO_4$ (LFP)/graphite battery lives using charge-discharge voltage data. Zhang et al. [24] identified battery degradation patterns from impedance spectroscopy using gaussian process machine learning models. Ding et al. [25] introduced a machine learning method for the improvement of the efficiency of membrane electrode assembly design and experiment. Such data-driven methods focus on the relationships among the input and output features, a key part of data-driven battery state estimation is the extraction of degradation features [26-28]. It has been proven that the relaxation process including the relaxation voltage value at a specific time and the voltage curve during a period shows a relationship with the battery SoH [11,12,16-18]. To our best knowledge, the relaxation process curve of the battery has not yet been studied with machine learning methods for large scale data. The relaxation process after a battery is fully charged is easy to obtain during EV operation, as there is no need of additional devices and controls consuming resources. Herein, an approach based on features extracted from the battery voltage relaxation is proposed, which focuses on short-term battery capacity estimation without any previous cycling information for on-board implementation. Four machine learning methods, i.e., ElasticNet [29], XGBoost [30], Support Vector Regression (SVR) [31], and Deep Neural Network (DNN) [32], using large datasets from two kinds of 18650-type commercial lithium-ion batteries, are employed in this study. Six statistical features, i.e., variance (VAR), skewness (SKE), maxima (MAX), minima (MIN), mean (MEAN), and kurtosis (KUR), extracted from the voltage relaxation curve together with the following discharge capacity are treated as one data unit. 27,330 data units from batteries with $LiNi_{0.86}Co_{0.11}Al_{0.03}O_2$ cathode (NCA battery) cycled at different temperature and current rates are used for model training and test. The root-mean-square errors (RMSEs) under different feature combinations are compared. It is found that all machine learning methods reach below 2.2% test RMSE by using three features (VAR, SKE, MAX) as input. XGBoost and SVR methods show the best performance with 1.1% test RMSE for NCA batteries. 30,312 data units from batteries with $LiNi_{0.83}Co_{0.11}Mn_{0.07}O_2$ cathode (NCM battery) are used to validate the model effectiveness. The DNN method obtains 1.8% RMSE, presenting the best estimation for the NCM battery without changing model weights.

Results

Data generation

Commercial 18650-type batteries are cycled in a temperature-controlled chamber with different charge current rates. Two battery brands with 3500 mAh (1C) nominal capacity, i.e., NCA battery and NCM battery, are selected for the dataset generation. The battery specifications are listed in Table.S1. Long-term cycling at three defined temperatures is conducted on 66 NCA cells and 55 NCM cells with a summary of cycling conditions provided in Table 1. Temperatures chosen are 25 °C, 35 °C and 45 °C. Three charge current rates (0.25C, 0.5C, 1C) are used for the NCA cells at 25 ºC. All discharge rates are set to 1C. The number of cells assigned to each cycling condition in Table 1 is aimed to obtain a dataset covering the variation between cells. One data unit

comprises a relaxation voltage curve after fully charged with the following discharge capacity. Each relaxation voltage curve is transformed into six statistical features, i.e., VAR, SKE, MAX, MIN, MEAN, and KUR. The datasets are collected for NCA and NCM cells, respectively. The NCA cell dataset with a total of 27,330 data units is used for model training and model test. The NCM cell dataset with 30,312 data units is used for assessing generalizability of the model.

Table 1 Cycling conditions for NCA and NCM batteries

Cell type	Cycling temperature (°C)	Charge current rate (C)	Number of cells	Number of data units
NCA battery 2.65 - 4.2V	25	0.25	7	1,825
		0.5	19	3,557
		1	9	373
	35	0.5	3	1,400
	45		28	20,175
NCM battery 2.5 - 4.2V	25		23	7,054
	35		4	4,716
	45		28	18,542

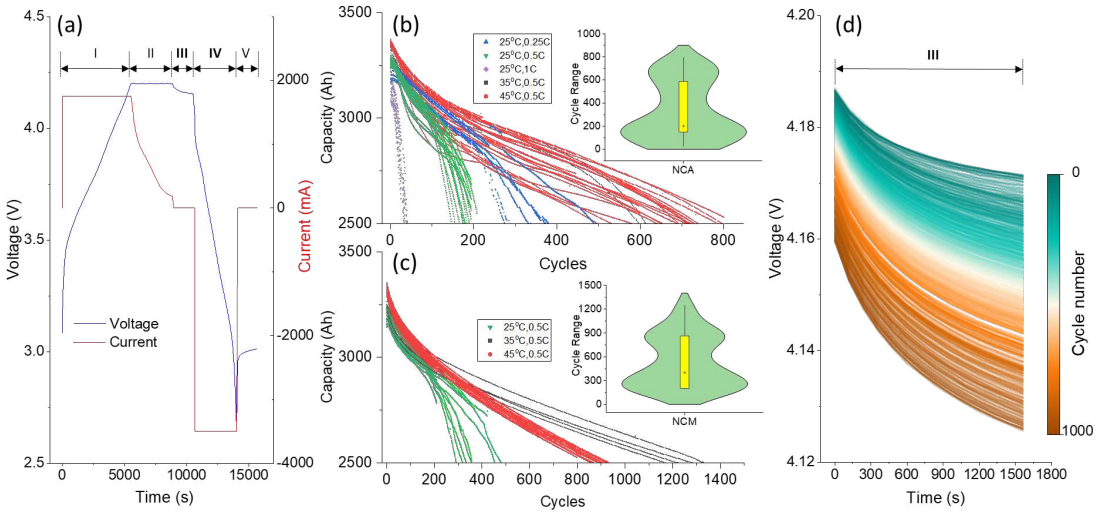

Fig.1 Battery cycling data. **a**, voltage and current profile in one cycle. **b,** NCA battery discharge capacity versus cycle number, with a violin plot of the cycle numbers at 2500 mAh. **c**, NCM battery discharge capacity versus cycle number, with a violin plot of the cycle numbers at 2500 mAh. **d**, schematic plot of relaxation voltage change (region III) while cycling for one NCA cell.

Voltage and current are the basic data recorded in this experiment, which includes charging, discharging, and relaxation processes. The cell cycling is performed with constant current (CC) charging to 4.2 V in a current rate ranging from 0.25 C to 1 C, followed by a constant voltage (CV) charging step at 4.2 V until a current corresponding to 0.05 C is reached. Constant current is then employed for the discharge at 3500 mA

(1C) to 2.65 V for the NCA cells and 2.5 V for the NCM cells, respectively. One complete cycling curve using 0.5 C charging rate for the NCA cell is shown in Fig.1a, which includes five processes, i.e. (I) CC charging, (II) CV charging, (III) relaxation after charging, (IV) CC discharging, and (V) relaxation after discharging. The CC discharging capacity is treated as the battery residual capacity during cycling. The relaxation time between the CV charging and CC discharging is 30 minutes with a sampling time of 120 s. A Biologic BCS potentiostat is used for the cell cycling in a temperature chamber. A dataset in a range of capacity fade to 2500 mAh (around 71% of nominal capacity) is generated. The battery capacity as a function of cycle number for the NCA cells is shown in Fig.1b. The cycle number is ranging from 50 to 800 in the 100% - 71% capacity window. It is evident that both, charging current and temperature have a strong influence on the capacity decay, and the battery capacity shows significant variance as depicted in the embedded violin plot, indicating the degradation distribution of the cycled cells. The worst scenario is the one with cells cycled at 1C at 25 ºC (CY25-1C), only 50 cycles can be obtained until the cells reach 71% of the nominal capacity. 71% capacity is reached after 125 and 600 cycles at 25 °C and 35 °C respectively, for cells charged with 0.5 C (CY25-0.5C, and CY35-0.5C). 71% capacity is reached after 250 cycles at 25 °C with 0.25 C charging current (CY25-0.25C) and in a range of 500 to 800 cycles at 45 °C with 0.5C charging current (CY45-0.5C). The cycling data of the NCM cells are shown in Fig. 1c. A fatigue down to 71% residual capacity is found between 250 and 500 cycles (25 °C), 1250 and 1500 cycles (35 °C) and around 1000 cycles at 45 °C cycling temperature. The capacity fade results indicate that increasing the temperature to 35 °C and 45 °C has a beneficial effect on the capacity retention, and that the charging current is at the limit of what the cells can handle. The major capacity loss is from loss of active lithium and is most likely consumed in the SEI formation as proven by in-situ neutron powder diffraction along with electrochemical analysis in our previous work [33]. Nevertheless, as this study aims to estimate the battery capacity based on data-driven methods, the relaxation process after fully charging is taken for feature extraction as the relaxation process is easily obtained in battery real use conditions. A schematic plot of relaxation voltage change against cycle number is presented in Fig.1d, which shows a declining trend with increasing cycle number. Summary statistics are proven to be effective to illustrate the shape and position change of the voltage curve numerically [23], thus, each voltage relaxation curve is converted to six statistical features, i.e., VAR, SKE, MAX, MIN, MEAN, and KUR, as displayed in Fig.2. The mathematical description of the six features is depicted in Table S2.

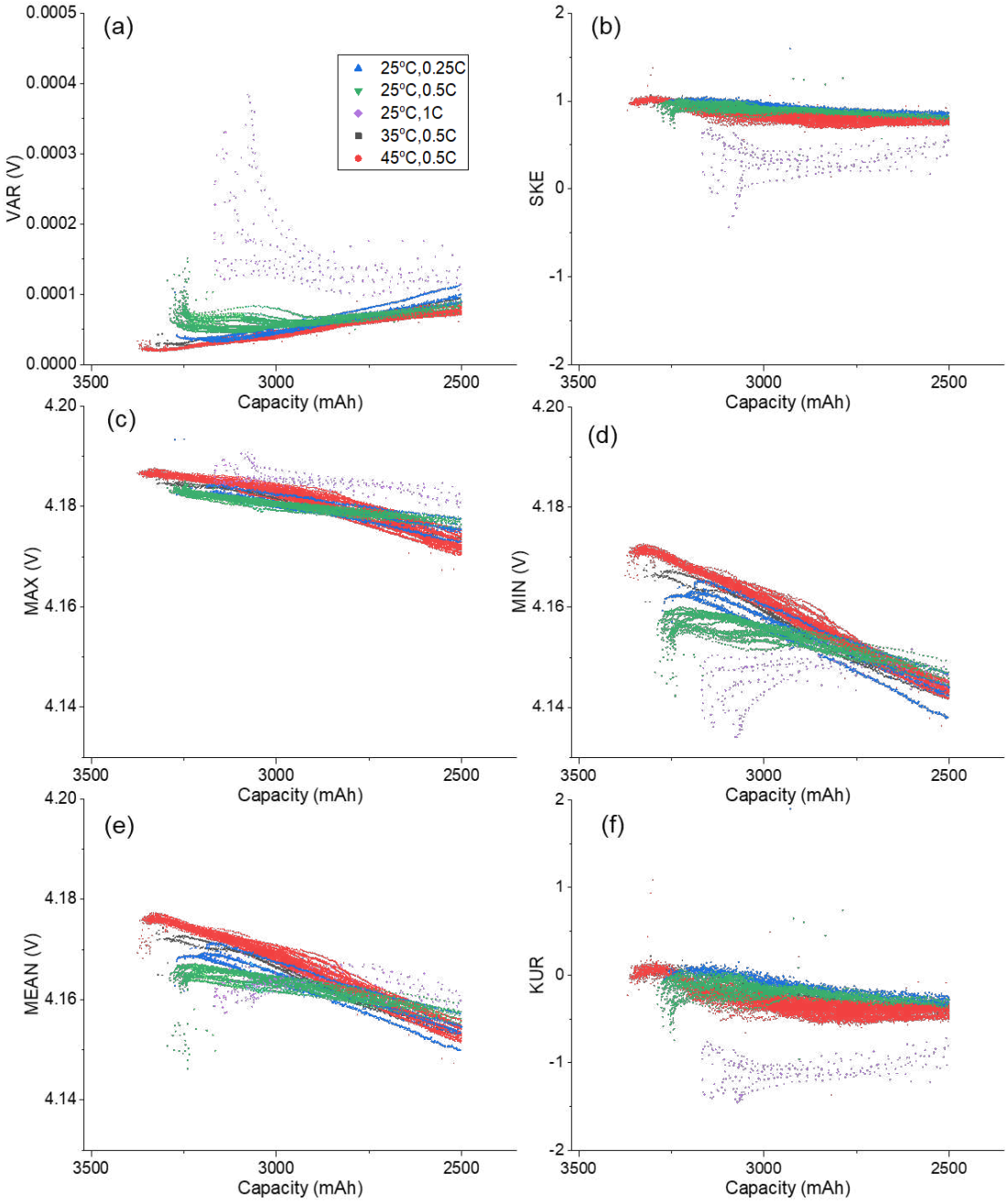

Fig.2 Extracted features from the voltage relaxation curves as a function of battery capacity for NCA cells. (**a**) VAR, (**b**) SKE, (**c**) MAX, (**d**) MIN, (**e**) MEAN, and (**f**) KUR.

The relationship between battery capacity and the corresponding features are dependent on the cycling conditions, from Fig.2, it is difficult to describe the relationships only by linear functions. The VAR in Fig.2a represents the distribution of the data, a decrease of VAR against capacity fade means that the relaxation voltages show a sharper distribution with increasing cycle number, and vice versa. Both SKE and KUR are normalized using VAR, they are used to describe the shape of the corresponding voltage curve. The SKE in Fig.2b is positive for almost all cycling conditions, indicating that more than half of the sampled voltage data are below the average voltage (MEAN), and which corresponds to the shape of the relaxation voltage

curve, i.e., the relaxation voltage drops fast at first and then gradually slows down. The MAX in Fig.2c presents a monotonous decrease of the maximum voltage with capacity drop for all cycling conditions. The MIN and MEAN first increase and then decrease versus the capacity reduction as displayed in Fig.2d and Fig.2e, respectively. The KUR shown in Fig. 2f is the excess kurtosis obtained from the kurtosis of the raw data minus the kurtosis of a normal distribution. The excess kurtosis is negative for all cycling conditions, meaning that the distribution of the relaxation voltage is flat.

Based on the extracted features from the relaxation voltage curve after charging, data-driven methods are used for battery capacity estimation. Owing to the difference in cathode materials and the manufacturing process for cycled batteries, a standard normalization for battery features and capacity is performed. Four machine learning algorithms, ElasticNet, XGboost, SVR, and DNN, are used in this study. ElasticNet [29] is a multivariate linear algorithm. XGBoost [30] is developed by improving ensemble algorithms including gradient boosting decision trees (GBDTs) and regularized greedy forest (RGF) algorithms [34]. SVR [31] is a kernel-based learning method which is used on data classification and clustering, regression estimation, and function approximation. DNN [32] is an artificial neural network with multiple layers and maps the nonlinearity between inputs and outputs. The hyperparameters of each algorithm are available in Table S3, and the results are compared in Fig.3.

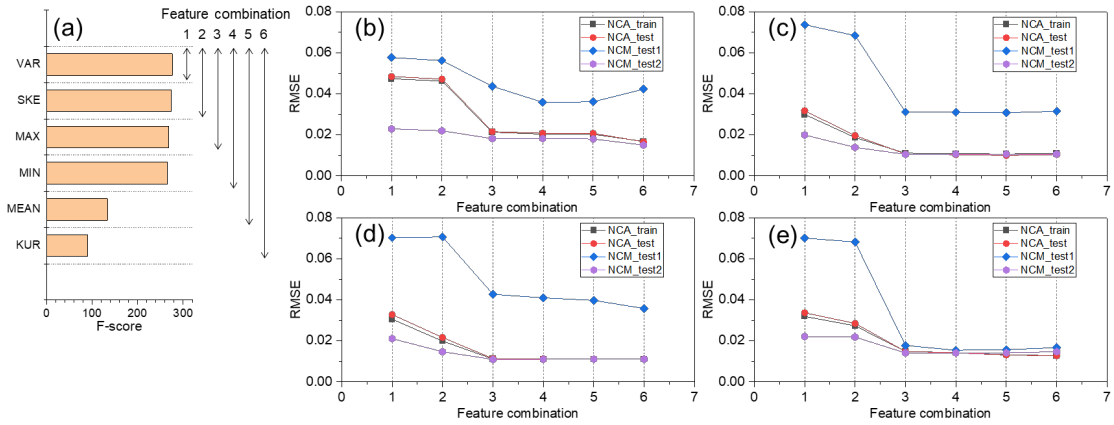

Fig.3 Capacity estimation results for cycled batteries. **a**, feature importance given by the F-score in XGBoost algorithm. **b-e**, RMSE change under different feature combinations by different estimation methods, ElasticNet (**b**), XGboost (**c**), SVR (**d**), and DNN (**e**). NCA_train is the RMSE of NCA battery model training, NCA_test presents the RMSE of the model test of NCA battery. NCM_test1 means that all NCM battery data are used in the model trained by NCA battery data without changing the model weights. NCM_test2 means the test results by modifying the model weights to adapt the self-characteristics of the NCM battery.

Feature reduction

XGBoost provides an assessment of the relevance of the individual input features. The importance of a feature in the XGBoost method is evaluated by the F-score [35]. The F-score is commonly adopted to measure the ability of features to distinguish between two categories and has been recognized as one of the best criteria to accomplish feature selection. The larger the F-score, the stronger is the discrimination ability of the

corresponding feature. The F-score results are presented in Figure 3a. It shows that VAR has the highest priority, followed by SKE and MAX. Correlation analysis is conducted to check the linear relationships between features. Strong correlations (close to ±1) indicate that these two features contain redundant information, meaning that a feature reduction is possible to reduce the model complexity. As shown in Table S4, the VAR does not show strong correlations with any other feature, indicating its discrimination ability. The SKE and KUR show a 0.96 correlation coefficient, likewise, MAX, MEAN, and MIN present a strong positive linear correlation (> 0.90). Therefore, the MAX is recommended as the representative feature. The results of the correlation analysis are consistent with the feature importance ranking from the F-score. Thus, feature reduction is conducted by using different feature combinations according to the F-score list. The feature combination 1 in Fig.3a means only the VAR is used in the model, 2 means the top two features (VAR and SKE), and so on. As two types of batteries are cycled for this study, the NCA cell dataset is used for the model building, and the NCM cell dataset is used for model verification. The capacity estimation results are summarized in Fig. 3b-3e.

Capacity estimation

For model training and test, firstly, the NCA data are randomly split into training set and test set in a 4:1 ratio. In the model training process, the K-fold Cross Validation with K=5 is used to determine the hyperparameters of the models. Further model tests are conducted using the NCM battery dataset. The results are summarized in Fig. 3b-3e, in which NCA_train is the RMSE of NCA battery model training, NCA_test presents the RMSE of the model test of NCA battery. To assess the generalizability of the model, two model verification strategies are conducted for the NCM battery data. NCM_test1 means that all NCM battery data are used in the model trained by NCA battery data without changing the model weights. To check the universality of methods, the NCM battery data are also split for training and test to adapt the self-characteristics of the NCM battery, NCM_test2 means the test results by modifying the model weights. The NCA_train and NCA_test RMSE curves in Fig.3b - 3e are close to each other, indicating the effectiveness of data splitting. The curves show that RMSE gradually decreases for all models, and the accuracy improvement is no longer obvious after using three features (VAR, SKE, and MAX), which agrees with the discussion in the feature reduction part as mentioned above. By comparing the RMSE with three features in table 2, it can be concluded that the RMSE of XGBoost and SVR reaches to 1.1%, showing better performance than the other two methods. The estimated capacity versus real capacity under three features is illustrated in Fig.4 and Fig.S1.

Table 2 RMSE of battery capacity estimation with three features (VAR, SKE, and MAX) by different estimation methods

	ElasticNet	XGboost	SVR	DNN
NCA_train	0.021	0.011	0.011	0.015
NCA_test	0.022	0.011	0.011	0.015
NCM_test1	0.044	0.031	0.043	0.018
NCM_test2	0.018	0.011	0.011	0.014

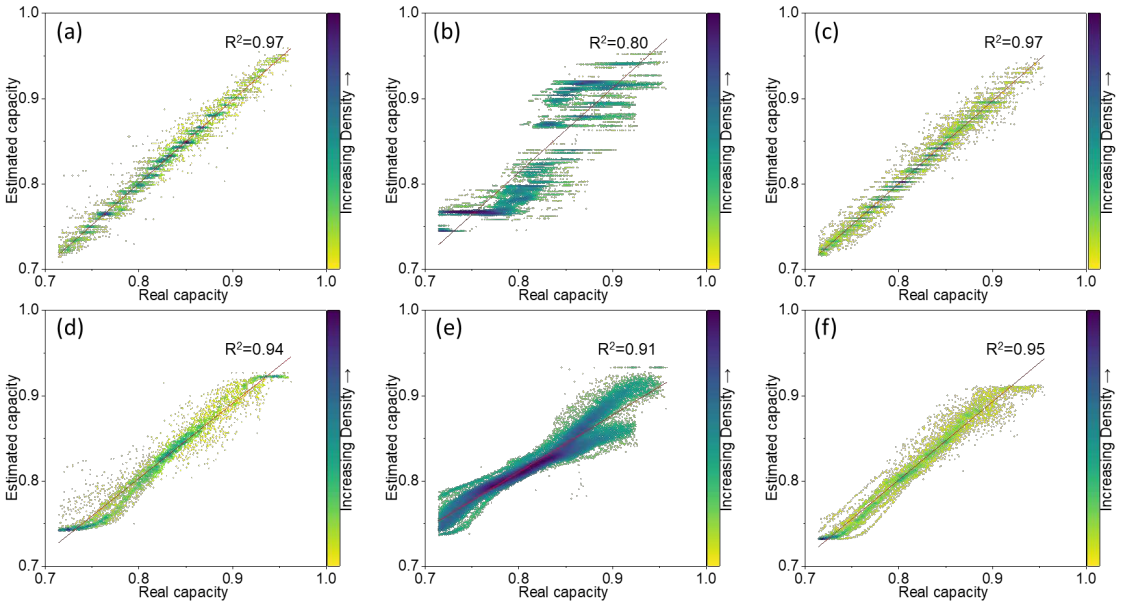

Fig.4 Test results of estimated capacity versus real capacity. NCA_test (**a**), NCM_test1 (**b**), and NCM_test2 (**c**) of XGBoost method. NCA_test (**d**), NCM_test1 (**e**), and NCM_test2 (**f**) of DNN method.

Verification of approach generalizability

The estimation results (NCA_test) by the XGBoost method are presented in Fig.4a. The coefficient of determination (R^2) is 0.97, showing that the model accurately estimates the battery capacity for the cycled NCA cells. The model trained by the NCA battery dataset without changing model weights is verified directly using 30,312 data units extracted from the NCM batteries. The XGBoost obtains about 3.1% test RMSE on NCM battery (NCM_test1) as presented Fig.4b. When the model is re-trained with the NCM battery dataset, the XGBoost method reaches a 1.1% test RMSE (NCM_test2), and the corresponding capacity results with R^2=0.97 are shown in Fig.4c. It can be found that the NCM_test1 RMSE (1.8%) of DNN is similar to NCA_test RMSE (1.5%) if the feature combination $\geqslant$ 3, as presented in Fig.3e. The corresponding capacity results are displayed in Fig.4e with R^2=0.91 and in Fig.4d with R^2=0.94, respectively. An estimation improvement is observed in Fig.4f, in which RMSE of the NCM_test2 is 1.4% and the R^2 is equal to 0.95. In summary, DNN gives better accuracy of capacity estimation if the model weights are not adapted, and the estimation accuracy is independent of the datasets. The method verification indicates that the proposed approach using the relaxation voltage curve can accurately estimate the battery capacity, and nonlinear methods are suggested to improve the estimation accuracy.

Discussion

Accurate identification of lithium-ion battery capacity facilitates the accurate estimation of driving range which is a primary concern for EVs. An approach without requiring information from previous cycling to estimate battery capacity is proposed. The proposed approach uses three statistical features (VAR, SKE, and MAX) extracted from the voltage relaxation curve as input. Four machine learning methods, i.e., ElasticNet, XGBoost, SVR, and DNN are applied on NCA and NCM batteries to

establish a suitable model and for the approach verification. The XGBoost, and SVR methods show good predictive ability with 1.1% RMSE for the NCA battery. The XGBoost obtains about 3.1% RMSE for the method verification on NCM batteries without changing the model weights, and it goes down to 1.1% test RMSE by modifying the model weights. The DNN method presents results within 1.5% RMSE for the NCA and NCM battery if the model weights are changed, and 1.8% RMSE for the NCM battery with the same model weights as trained by the NCA battery dataset, indicating the applicability of the proposed capacity estimation approach. This work promotes the development of using data-driven methods for battery SoH estimation in EVs.

Methods

Cell selection and cycling

Commercially available lithium-ion batteries with a nominal capacity of 3500mAh (INR18650-35E and INR18650-MJ1) have been tested. Inductively coupled plasma optical emission spectrometry (ICP-OES) shows that the composition of the cathode of a fresh cell in the discharged state is $LiNi_{0.86}Co_{0.11}Al_{0.03}O_2$ for the INR18650-35E batteries and $Li(Ni_{0.83}Co_{0.11}Mn_{0.07})O_2$ for the INR18650-MJ1 batteries. The anode composition for both cell types is determined by a Carbon Hydrogen Nitrogen (CHN) Analyzer to have roughly 97 wt% C and 2 wt% Si as well as traces of H, N and S. The INR18650-35E battery is named as NCA battery, and the INR18650-MJ1 is named as NCM battery according to their cathode material. A biologic BCS potentiostat is employed for the cell cycling and the measurements are conducted in a climate chamber. Long term cycling is conducted on 66 NCA cells and 55 NCM cells with a summary of cycling conditions as provided in Table 1.

ElasticNet method

The ElasticNet algorithm is proposed by Zou et al. [29], which is a regularized regression method that linearly combines the L_1 and L_2 penalties of the lasso and ridge methods. ElasticNet is an extension of ordinary least square (OLS) regression. In OLS regression, given d features $x_{i1}, \ldots, x_{id}$, the response y_i is predicted by:

$$\hat{y}_i = \beta_0 + \sum_{j=1}^{d} \beta_j x_{ij} \tag{1}$$

A model fitting procedure produces the parameter vector $\hat{\beta} = (\hat{\beta}_0, \ldots, \hat{\beta}_d)$.

For the data set having n observations with p features, let $y = (y_1, \ldots, y_n)^T$, $X = \begin{bmatrix} x_{11} & \cdots & x_{1d} \\ \cdots & x_{ij} & \cdots \\ x_{n1} & \cdots & x_{nd} \end{bmatrix}$.

The ElasticNet loss function is defined as：

$$L(\lambda_1, \lambda_2, \beta) = \|y - X\beta\|_2^2 + \lambda_2\|\beta\|_2^2 + \lambda_1\|\beta\|_1 \tag{2}$$

If we set $\alpha = \lambda_2/(\lambda_1 + \lambda_2)$, the optimized parameters vector is obtained by:

$$\hat{\beta} = \mathrm{argmin}_\beta L(\alpha, \beta) = \|y - X\beta\|_2^2 + \alpha\|\beta\|_2^2 + (1-\alpha)\|\beta\|_1 \tag{3}$$

where $\alpha\|\beta\|_2^2 + (1-\alpha)\|\beta\|_1$ is called the ElasticNet penalty, which is a convex

combination of the lasso and ridge penalty.

XGBoost method

The XGBoost method [30] is a scalable end-to-end tree boosting system designed to be highly efficient, flexible, and portable. It implements machine learning algorithms in the Gradient Boosting framework. Compared with multiple linear regression, XGBoost has the advantage of being able to handle nonlinear relationships. The tree $f(x)$ is defined as:

$$f_t(x) = \omega_{q(x)}, (q: \mathbb{R}^d \to \{1,2,\dots,T\}, \omega \in \mathbb{R}^T) \tag{4}$$

where t represents a tree, q represents the structure of each tree that maps an example to the corresponding leaf index. T is the number of leaves in the tree. Each f_t corresponds to an independent tree structure q and leaf weights ω(output of a tree).

The objective function is defined as:

$$obj^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{i=1}^{t} \Omega(f_i) \tag{5}$$

where l is a differentiable convex loss function that measures the difference between the prediction $\hat{y}$ and the target y_i. The second term Ω penalizes the complexity of the model, which helps to smooth the final learnt weights to avoid over-fitting.

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \tag{6}$$

where ω_j is the weight of the j^{th} leaf node. γ and λ are the coefficients for penalty term Ω.

Using the second-order Taylor's formula, the objective function can be given as:

$$\begin{aligned} obj^{(t)} &= \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) + constant \\ &\approx \sum_{i=1}^{n} \left(l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right) + \Omega(f_t) + constant \end{aligned} \tag{7}$$

where x_i is the input of the sample, $g_i = \partial_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}^{(t-1)}\right)$ and $h_i = \partial^2_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}^{(t-1)}\right)$

After removing the constant, the objective function at step t becomes

$$\begin{aligned} obj^{(t)} &\approx \sum_{i=1}^{n} \left(g_i \omega_{q(x_i)} + \frac{1}{2} h_i \omega^2_{q(x_i)}\right) + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \\ &= \sum_{j=1}^{T} \left(G_j \omega_j + \frac{1}{2}(H_j + \lambda)\omega_j^2\right) + \gamma T \end{aligned} \tag{8}$$

Where $G_j = \sum_{i \in I_j} g_i, H_j = \sum_{i \in I_j} h_i,\ I_j = \{i | q(x_i = j)\}$ is instance set of leaf j.

the optimal weight ω_j^* of leaf j for a fixed structure $q(x)$ can be computed by:

$$\omega_j^* = -\frac{G_j}{H_j + \lambda} \tag{9}$$

The optimal loss is:

$$obj^* = -\frac{1}{2}\sum_{j=1}^{T} \frac{\left(G_j\right)^2}{H_j + \lambda} + \gamma T \tag{10}$$

obj^* is a function of marking tree structure and measuring the quality of tree structure q. The smaller the value of obj^*, the better.

SVR method

SVR approach [31] is a kernel-based method which does not regress on the original input vector, but on its nonlinear expansion, which is mapped from a kernel function to a very high-dimensional feature space. Given a training set of data $\{(x_1, y_1), \ldots (x_n, y_n)\}$, where $x_i \subset R^d$ donates the input space of the sample, $y_i \subset R$ is the target value. $i = 1, \ldots, n,$ corresponds to the size of the training data.

The generic SVR estimating function takes the form

$$\hat{y}_i = \left(\omega \cdot \Phi(x_i)\right) + b \tag{11}$$

where $\omega \subset R^d, b \subset R,$ and $\Phi(x)$ is a nonlinear transformation from R^d to a high-dimensional space. The ω has the following expansion:

$$\omega = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*)\Phi(x_i) \tag{12}$$

where $\alpha_i \, and \, \alpha_i^*$ are the Lagrange multiplier. With the expression of the kernel function $k(x_i, x) = \Phi(x_i) \cdot \Phi(x)$, the SVR estimating function can be expressed as:

$$\hat{y}_i = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*)\, k(x_i, x) + b \tag{13}$$

The goal of SVR is to find the value of ω and b that minimizing the total loss

$$\min\left\{\frac{1}{2}\|\omega\|_2^2 + C\sum_{i=1}^{l} \ell_\epsilon(\hat{y}_i - y_i)\right\} \tag{14}$$

where C is a constant, and vector ℓ_ϵ is the loss function, the ϵ-insensitive loss function is used in this research:

$$\ell_\epsilon(\hat{y}_i - y_i) = \begin{cases} |\hat{y}_i - y_i| - \epsilon, & |\hat{y}_i - y_i| \geq \epsilon \\ 0, & |\hat{y}_i - y_i| < \epsilon \end{cases} \tag{15}$$

DNN method

A deep neural network (DNN) [32] is an artificial neural network with multiple layers as presented in Fig.S2. It can approximate the non-linear mapping between inputs and outputs. For a DNN with *m* hidden layers, the output is predicted by

$$\hat{y}_i = \varphi\left(v^T\varphi\left(W^{(m)}\varphi\left(\cdots W^{(2)}\varphi\left(W^{(1)}x_i\right)\right)\right)\right) \quad (16)$$

where x_i is the input of the model (i.e., three statistical features VAR, SKE, MAX), v and W are the weights of output layer and hidden layers, respectively, $\varphi(\cdot)$ is the activation function, which can introduce the nonlinearity of the model. For different layers in the DNN model, the activation function can be different. In our study, the DNN has five dense hidden layers and each with a *sigmoid* activation function. The output layer predicts the relative capacity using an *softplus* activation function, which ensures that the output is greater than 0 Dropout layers with a rate of 0.1 are also added between dense layers to avoid over-fitting. A summary of the NN model structure can be found in Table S3.

The goal of training a DNN model is to find the weights of v and W that minimizing the total loss

$$\min\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \quad (17)$$

where $\hat{y}_i$ is the predicted relative capacity generated by the model and y_i is the corresponding true relative capacity. The NN is trained with Keras and Adam algorithm is used as the optimizer to update the weights of the model.

Data availability

The data that support the findings of this study are available at *publication*, further information and request for resources should be directed to the contact, Dr. Michael Knapp and Dr. Haifeng Dai. Full access to data can be required for peer review.
review

Code availability

The data processing and the proposed method is performed in python and is available at *publication*. Full access to code can be required for peer review.

References

1 Bresser, D. *et al.* Perspectives of automotive battery R&D in China, Germany, Japan, and the USA. *J. Power Sources* **382**, 176-178, doi:10.1016/j.jpowsour.2018.02.039 (2018).

2 Harper, G. *et al.* Recycling lithium-ion batteries from electric vehicles. *Nature* **575**, 75-86 (2019).

3 Waag, W., Käbitz, S. & Sauer, D. U. Experimental investigation of the lithium-ion battery impedance characteristic at various conditions and aging states and its influence on the application. *Applied Energy* **102**, 885-897, doi:10.1016/j.apenergy.2012.09.030 (2013).

4 Xiong, R., Li, L. & Tian, J. Towards a smarter battery management system: A critical review on battery state of health monitoring methods. *J. Power Sources* **405**, 18-29 (2018).

5 Tang, X. *et al.* A fast estimation algorithm for lithium-ion battery state of health. *J. Power Sources* **396**, 453-458 (2018).

6 Goh, T., Park, M., Seo, M., Kim, J. G. & Kim, S. W. Capacity estimation algorithm with a second-order differential voltage curve for Li-ion batteries with NMC cathodes. *Energy* **135**, 257-268 (2017).

7 Pei, P. *et al.* Capacity estimation for lithium-ion battery using experimental feature interval

approach. *Energy*, 117778 (2020).

8 Jiang, B., Dai, H. & Wei, X. Incremental capacity analysis based adaptive capacity estimation for lithium-ion battery considering charging condition. *Applied Energy* **269**, 115074 (2020).

9 Eddahech, A., Briat, O. & Vinassa, J.-M. Determination of lithium-ion battery state-of-health based on constant-voltage charge phase. *J. Power Sources* **258**, 218-227 (2014).

10 Baghdadi, I., Briat, O., Gyan, P. & Vinassa, J. M. State of health assessment for lithium batteries based on voltage–time relaxation measure. *Electrochim. Acta* **194**, 461-472 (2016).

11 Schindler, S., Bauer, M., Petzl, M. & Danzer, M. A. Voltage relaxation and impedance spectroscopy as in-operando methods for the detection of lithium plating on graphitic anodes in commercial lithium-ion cells. *J. Power Sources* **304**, 170-180 (2016).

12 von Lüders, C. *et al.* Lithium plating in lithium-ion batteries investigated by voltage relaxation and in situ neutron diffraction. *J. Power Sources* **342**, 17-23 (2017).

13 Pei, L., Wang, T., Lu, R. & Zhu, C. Development of a voltage relaxation model for rapid open-circuit voltage prediction in lithium-ion batteries. *J. Power Sources* **253**, 412-418 (2014).

14 Li, A., Pelissier, S., Venet, P. & Gyan, P. Fast characterization method for modeling battery relaxation voltage. *Batteries* **2**, 7 (2016).

15 Deng, L., Shen, W., Wang, H. & Wang, S. A rest-time-based prognostic model for remaining useful life prediction of lithium-ion battery. *Neural Computing and Applications*, 1-12 (2020).

16 Qian, K. *et al.* State-of-health (SOH) evaluation on lithium-ion battery by simulating the voltage relaxation curves. *Electrochim. Acta* **303**, 183-191 (2019).

17 Attidekou, P. S., Wang, C., Armstrong, M., Lambert, S. M. & Christensen, P. A. A New Time Constant Approach to Online Capacity Monitoring and Lifetime Prediction of Lithium Ion Batteries for Electric Vehicles (EV). *J. Electrochem. Soc.* **164**, A1792 (2017).

18 Fang, Q., Wei, X., Lu, T., Dai, H. & Zhu, J. A State of Health Estimation Method for Lithium-Ion Batteries Based on Voltage Relaxation Model. *Energies* **12**, 1349 (2019).

19 Li, W. *et al.* Digital twin for battery systems: Cloud battery management system with online state-of-charge and state-of-health estimation. *Journal of Energy Storage* **30**, 101557 (2020).

20 Liu, K., Shang, Y., Ouyang, Q. & Widanage, W. D. A data-driven approach with uncertainty quantification for predicting future capacities and remaining useful life of lithium-ion battery. *IEEE Transactions on Industrial Electronics* (2020).

21 Li, W. *et al.* Online capacity estimation of lithium-ion batteries with deep long short-term memory networks. *J. Power Sources* **482**, 228863 (2021).

22 Hu, X., Xu, L., Lin, X. & Pecht, M. Battery lifetime prognostics. *Joule* **4**, 310-346 (2020).

23 Severson, K. A. *et al.* Data-driven prediction of battery cycle life before capacity degradation. *Nature Energy* **4**, 383-391 (2019).

24 Zhang, Y. *et al.* Identifying degradation patterns of lithium ion batteries from impedance spectroscopy using machine learning. *Nature communications* **11**, 1-6 (2020).

25 Ding, R. *et al.* Designing AI‐Aided Analysis and Prediction Models for Nonprecious Metal Electrocatalyst‐Based Proton‐Exchange Membrane Fuel Cells. *Angew. Chem. Int. Ed.* **59**, 19175-19183 (2020).

26 Lin, C., Cabrera, J., Denis, Y., Yang, F. & Tsui, K. SOH Estimation and SOC Recalibration of Lithium-Ion Battery with Incremental Capacity Analysis & Cubic Smoothing Spline. *J. Electrochem. Soc.* **167**, 090537 (2020).

27 Tagade, P. *et al.* Deep Gaussian process regression for lithium-ion battery health prognosis and

degradation mode diagnosis. *J. Power Sources* **445**, 227281 (2020).

28 Chen, K. *et al.* Practical failure recognition model of lithium-ion batteries based on partial charging process. *Energy* **138**, 1199-1208 (2017).

29 Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. Roy. Stat. Soc. Ser. B. (Stat. Method.)* **67**, 301-320 (2005).

30 Chen, T. & Guestrin, C. in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining.* 785-794.

31 Awad, M. & Khanna, R. in *Efficient learning machines* 67-80 (Springer, 2015).

32 Tian, J., Xiong, R., Shen, W., Lu, J. & Yang, X.-G. Deep neural network battery charging curve prediction using 30 points collected in 10 min. *Joule* **5**, 1521-1534 (2021).

33 Sørensen, D. R. *et al.* Fatigue in High-Energy Commercial Li Batteries while Cycling at Standard Conditions: An In Situ Neutron Powder Diffraction Study. *ACS Applied Energy Materials* **3**, 6611-6622 (2020).

34 Liu, K. *et al.* Feature analyses and modelling of lithium-ion batteries manufacturing based on random forest classification. *IEEE/ASME Transactions on Mechatronics* (2021).

35 Guyon, I., Gunn, S., Nikravesh, M. & Zadeh, L. A. *Feature extraction: foundations and applications*. Vol. 207 (Springer, 2008).

Acknowledgements

This work contributes to the research performed at CELEST (Center for Electrochemical Energy Storage Ulm-Karlsruhe) and supported in the frame of the Alexander von Humboldt Postdoctoral Research Program. The authors would like to thank the foundations of the National Natural Science Foundation of China (NSFC, Grant No. U1764256 and 51677136).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementary20210801.docx